МИЭТ НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ УНИВЕРСИТЕТ

Кафедра «Информационная безопасность

XXXI научно-практическая конференция «Комплексная защита информации»
6-7 октября 2026 г. г. Москва, Россия

ОЦЕНКА РАЗБОРЧИВОСТИ РЕЧИ В УСЛОВИЯХ ПОМЕХ С ИСПОЛЬЗОВАНИЕМ ТЕХНОЛОГИЙ ИСКУССТВЕННОГО ИНТЕЛЛЕКТА

Хорев Анатолий Анатольевич

профессор, доктор технических наук

2026 год

Сравнительный анализ систем автоматического распознавания речи

Модель САРР	Процент ошибок, рассчитанный по метрике ***WER*** *(%)*		
	Golos Farfield	Russian MCV-19	Russian LibriSpeech
XLS_R finetuned	15,7	8,0	16,1
Whisper-large-v3 (США)	16,6	5,5	9,5
FastConformer-RNNT	6,6	5,7	11,3
GigaAM-CTC-v3	4,3	3,1	5,5
GigaAM-RNNT-v3 (Россия)	3,9	2,7	5,5

$$\boldsymbol{WER} = \frac{(\text{пропуски} + \text{замены} + \text{вставки})\ \text{слов}}{\text{общее количество слов}} \times 100\%$$

Характеристики выбранных систем автоматического распознавания речи

Характеристика	GigaAM-RNNT-v3	WhisperX-large-v3
Тип системы	Сквозная потоковая модель распознавания	Программный конвейер с оффлайн моделью распознавания
Основная задача	Распознавание русской речи в реальном времени	Временно-точное распознавание длинных аудио
Архитектура ядра распознавания	Конформер-кодер + RNNT-декодер	Трансформер-кодер авторегрессионный декодер
Детектор речевой активности	внешний Silero ДРА	WhisperX ДРА
Предобучение	HuBERT-CTC. **700 тыс. час. русской речи**	Слабо-контролируемое обучение на 680 тыс. час размеченных и 5 млн неразмеченных многоязычных и многозадачных данных
Тонкая настройка	на 4 тыс. часах на русских корпусах Golos farfield, MCV и Russian LibriSpeech	Не производилось
Поддержка русского языка	**Имеется**	**Имеется**
Размер модели	Кодер 240M, общий размер ~240M, т.к. RNNT-декодер крошечный	Кодер-декодер Whisper-large-v3 – 1,55 млрд параметров
Требуемые вычислительные ресурсы	CPU/GPU среднего уровня	Мощный GPU
Открытость/лицензия	MIT лицензия, код и веса выложены	MIT лицензия, код и веса выложены
Важные особенности	Удаление сегментов с более 60% тишины при обучении	принудительное фонемное выравнивание

Методика проведения экспериментальных исследований

1) Производилась запись исходных тестовых речевых сигналов (таблиц слов и фраз) проводилась бригадой из 6–ти дикторов (4 мужчины и 2 женщины), не имеющих дефектов речи.
2) Были сформированы записи зашумленных речевых сигналов при помощи программы Adobe Audition 1.5 (вид шума – «белым» шум, шумовая «речеподобная» помеха, помеха типа «речевой хор») с отношением сигнал/шум (*SNR*) от – 20 дБ до + 10 дБ.
3) Аудиофайлы с записями зашумленных речевых сигналов загружались в САРР.
4) САРР проводила распознавание слов или фраз.
5) Результат распознавания сравнивался с эталоном.
7) По итогу сравнения рассчитывалось количество правильно распознанных слов/фраз по формуле:

$$W_{ik} = N_{ik}/N,$$

где W_{ik}– словесная (фразовая) разборчивость аудиофайла с записью слов (фраз) i–й таблицы, определенная k–м аудитором;

N_{ik} – количество правильно распознанных k–м аудитором слов/фраз в аудиофайле с записью слов/фраз i–й таблицы (слово считалось правильно понятым АСР, если после нормализации оно совпадало со словом из эталонного текста; Фраза считалась правильно);

N – общее количество слов/фраз в одном аудиофайле (N= 50)

МИЭТ НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ УНИВЕРСИТЕТ

Результаты оценки словесной разборчивости речи, распознанной САРР при высоких отношениях сигнал/шум

САРР «GigaAM-RNNT-v3»

Диктор	1 таблица	2 таблица	3 таблица
1	62	78	94
2	62	68	78
3	72	74	46
4	58	46	44
5	22	56	24
6	76	54	54
	58,67	62,67	56,67
	59,3 %		

Диктор	1 таблица	2 таблица	3 таблица
1	100	100	98
2	94	94	96
3	90	100	94
4	94	84	88
5	94	90	100
6	98	86	96
	95	92,3	95,3
	94,2 %		

САРР «WhisperX-large-v3»

Диктор	1 таблица	2 таблица	3 таблица	4 таблица	5 таблица
1	100	100	100	100	100
2	100	100	100	100	98
3	100	100	100	100	100
4	100	100	100	100	100
5	100	98	100	98	96
6	100	100	98	100	98
	100	99,7	99,7	99,7	98,7
	99,5 %				

МИЭТ НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ УНИВЕРСИТЕТ

Результаты оценки словесной разборчивости зашумленной речи для «белого» шума

САРР «GigaAM-RNNT-v3» (диктор 1, таблица 1)

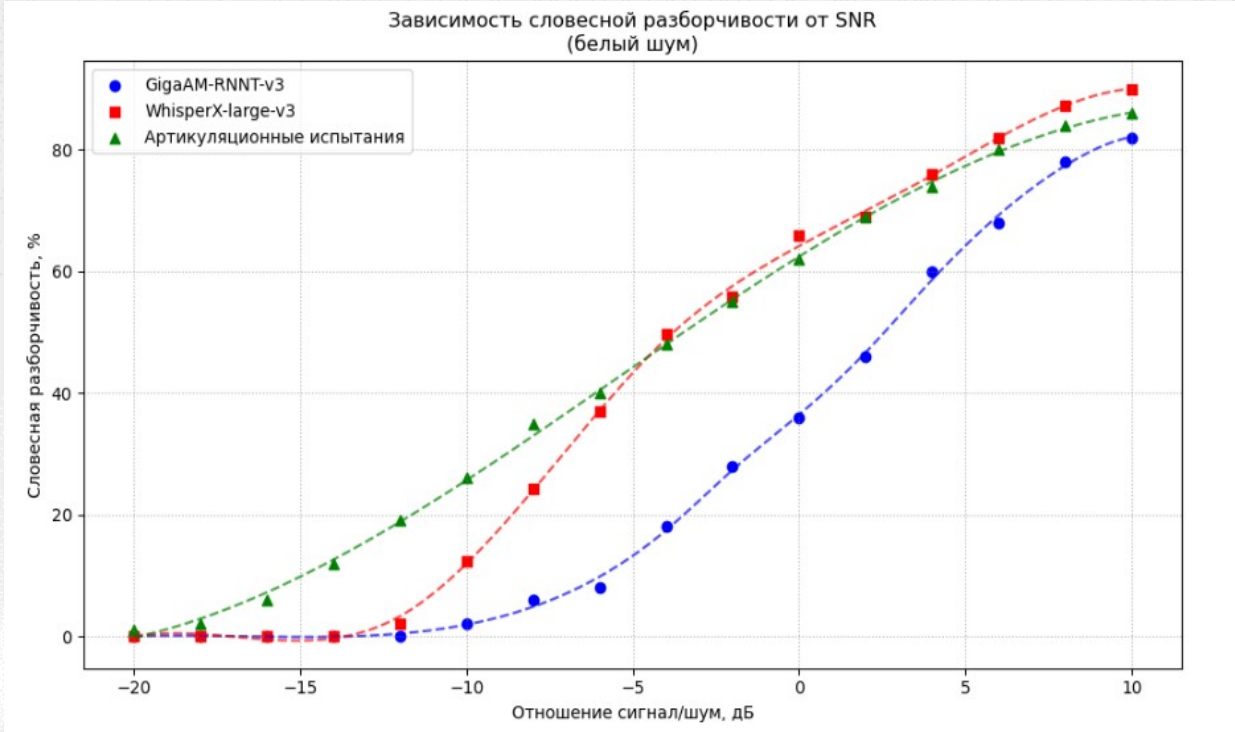
SNR, дБ	
–20	0
–18	0
–16	0
–14	0
–12	0
–10	2
–8	6
–6	8
–4	18
–2	28
0	36
2	46
4	60
6	68
8	78
10	82

САРР «WhisperX –large–v3» (6 дикторов, 6 таблиц)

SNR, дБ	Диктор 1	Диктор 2	Диктор 3	Диктор 4	Диктор 5	Диктор 6	
–20	0	0	0	0	0	0	0
–18	0	0	0	0	0	0	0
–16	0	0	0	0	0	0	0
–14	0	0	0	0	0	0	0
–12	0	0	4	0	4	4	2
–10	0	16	16	26	4	12	12,3
–8	24	20	32	36	6	28	24,3
–6	30	30	48	46	12	56	37
–4	52	52	62	58	18	56	49,7
–2	54	60	68	60	28	64	55,8
0	60	76	74	66	38	80	65,8
2	68	76	80	70	40	80	69
4	72	82	90	–	48	88	76
6	88	90	90	78	56	90	82
8	94	90	94	82	70	94	87,3
10	96	90	94	88	76	96	90

Результаты оценки словесной разборчивости зашумленной речи для «белого» шума

SNR, дБ	«GigaAM-RNNT-v3»	«WhisperX-large-v3»	Артикуляционные испытания
–20	0	0	1
–18	0	0	2
–16	0	0	6
–14	0	0	12
–12	0	2	19
–10	2	12,3	26
–8	6	24,3	35
–6	8	37	40
–4	18	49,7	48
–2	28	55,8	55
0	36	65,8	62
2	46	69	69
4	60	76	74
6	68	82	80
8	78	87,3	84
10	82	90	86



Оценка словесной разборчивости зашумленной речи для «белого» шума методом артикуляционных испытаний с проводилась с привлечением 8 аудиторов

МИЭТ НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ УНИВЕРСИТЕТ

Результаты оценки фразовой разборчивости речи, распознанной САРР при высоких отношениях сигнал/шум

САРР «GigaAM-RNNT-v3»

Диктор	1 табл.	2 табл.	3 табл.	4 табл.	5 табл.
1	100	86	76	80	66*
2	98	98	98	96	90
3	94	94	98	86	96
4	98	88	84	82	90
5	-	94	92	82	84
6	90	96	100	94	96
	96	92,67	91.33	86,67	87
	90,7				

САРР «WhisperX-large–v3»

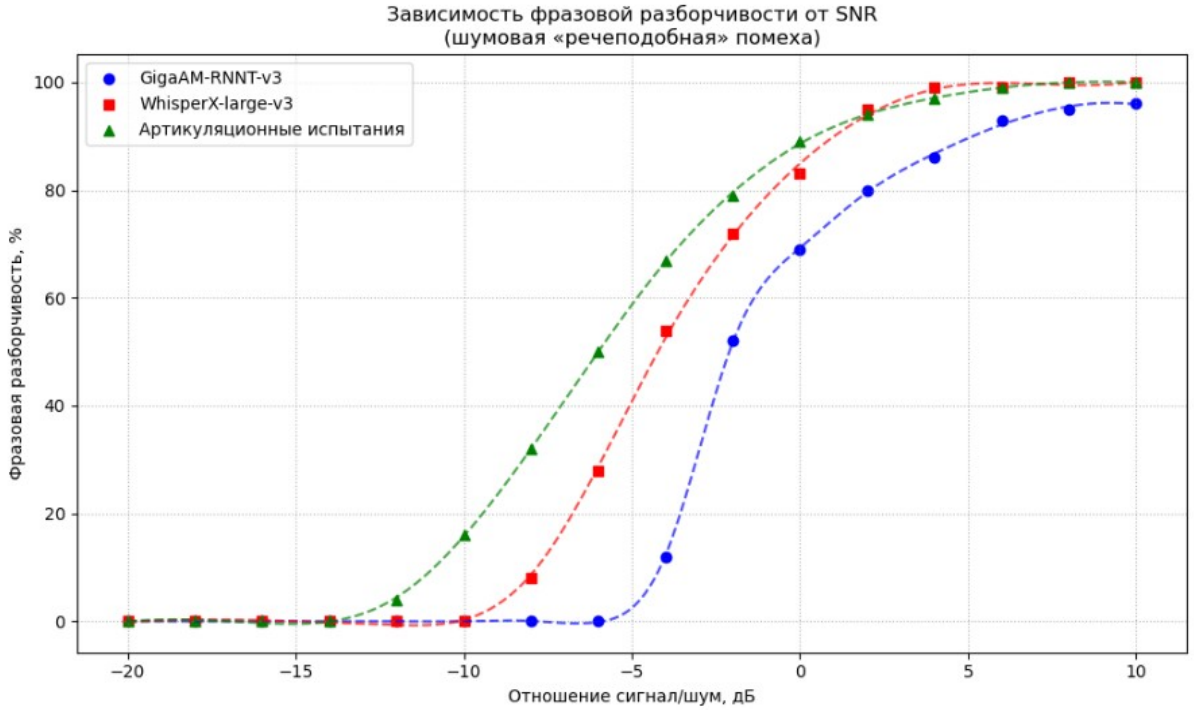
Диктор	1 табл.	2 табл.	3 табл.	4 табл.	5 табл.
1	100	100	100	100	98*
2	100	100	100	100	100
3	100	100	100	100	100
4	100	100	100	100	100
5	-	100	100	100	100
6	100	100	100	100	100
	100	100	100	100	99,7
	99,9				

Результаты оценки фразовой разборчивости зашумленной речи

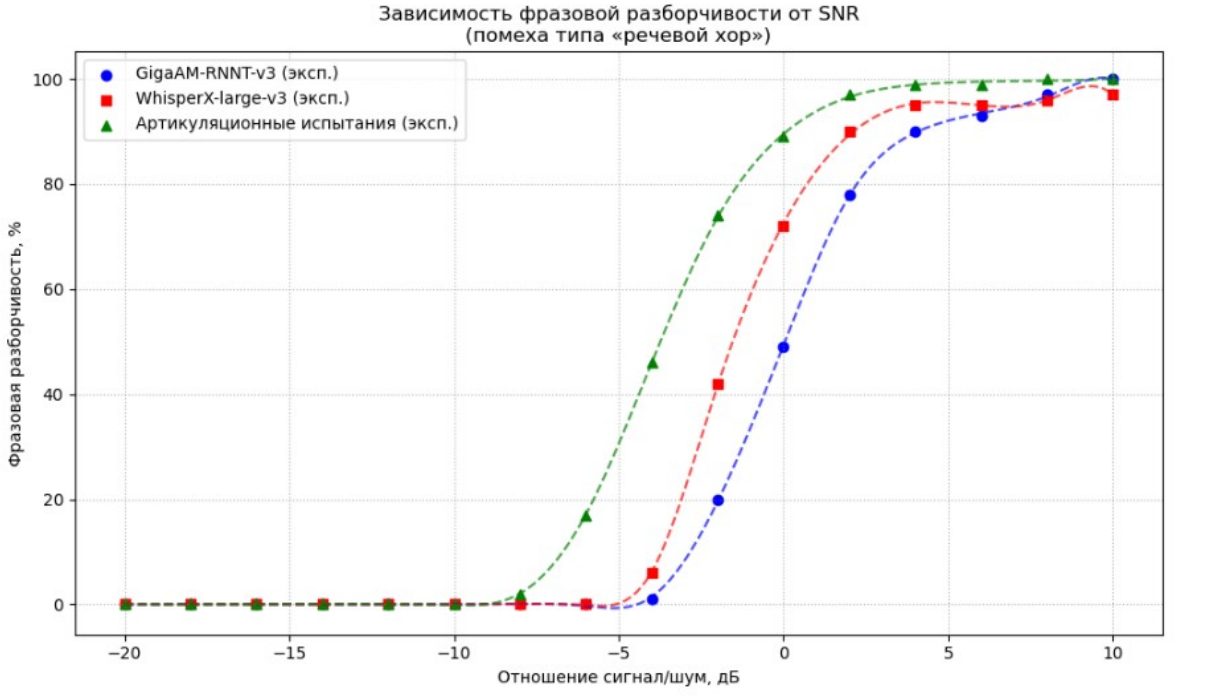
SNR, дБ	Шумовая «речеподобная» помеха			Помеха типа «речевой хор»		
	«GigaAM-RNNT-v3»	«WhisperX-large-v3»	Артикуля-ц. испытания	«GigaAM-RNNT-v3»	«WhisperX-large-v3»	Артикуляционные испытания
-20	0	0	0	0	0	0
-18	0	0	0	0	0	0
-16	0	0	0	0	0	0
-14	0	0	0	0	0	0
-12	0	0	4	0	0	0
-10	0	0	16	0	0	0
-8	0	8	32	0	0	2
-6	0	28	50	0	0	17
-4	12	54	67	1	6	46
-2	52	72	79	20	42	74
0	69	83	89	49	72	89
2	80	95	94	78	90	97
4	86	99	97	90	95	99
6	93	99	99	93	95	99
8	95	100	100	97	96	100
10	96	100	100	100	97	100

Результаты оценки фразовой разборчивости зашумленной речи

Шумовая «речеподобная» помеха



Речевая помеха типа «речевой хор»



Выводы

1) **При оценке словесной разборчивости зашумленных речевых сигналов:**

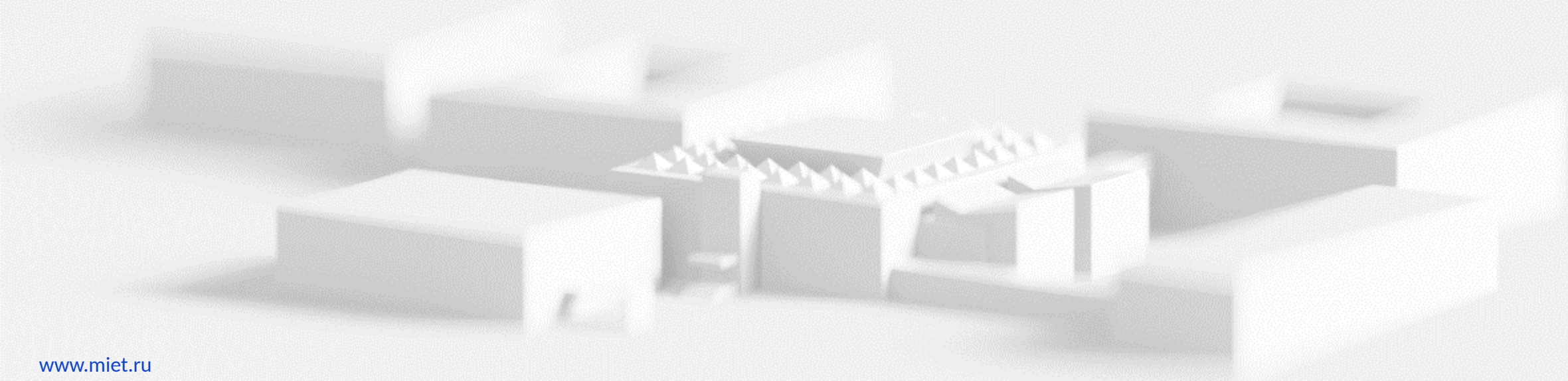
- при отношениях *SNR* < 6 дБ словесная разборчивость зашумленных речевых сигналов, полученная САРР «GigaAM-RNNT-v3», значительно ниже словесной разборчивости речи, полученной САРР «WhisperX-large-v3»;
- результаты оценки словесной разборчивости зашумленных речевых сигналов, полученные САРР «WhisperX-large–v3» , сопоставимы с результатами артикуляционных испытаний при *SNR* > - 6 дБ;
- при *SNR* ≤ - 8 дБ словесная разборчивость зашумленных речевых сигналов, полученная САРР «GigaAM-RNNT-v3», существенно ниже результатов артикуляционных испытаний;

2) При оценке фразовой разборчивости зашумленных речевых сигналов:

фразовая разборчивость зашумленных речевых сигналов, полученная САРР «GigaAM-RNNT-v3», значительно ниже разборчивости, полученной САРР «WhisperX-large-v3»;

а) для шумовой «речеподобной» помехи при отношении *SNR* < –4 дБ фразовая разборчивость зашумленных речевых сигналов, полученная САРР «WhisperX-large–v3», значительно ниже фразовой разборчивости речи, полученной по результатам артикуляционных испытаний;

б) для помехи типа «речевой хор» при отношении *SNR* < –2 дБ, фразовая разборчивость зашумленных речевых сигналов, полученная САРР «WhisperX-large–v3», значительно ниже фразовой разборчивости речи, полученной по результатам артикуляционных испытаний.

Таким образом, анализ полученных результатов экспериментальных исследований показал, что САРР «WhisperX-large-v3» может быть использована взамен артикуляционных испытаний: для оценки словесной разборчивости речи при использовании помехи типа «белый шум» - при отношениях SNR > - 6 дБ; для оценки фразовой разборчивости речи при использовании шумовой «речеподобной помехи» - при отношениях SNR > 0 дБ, а при использовании помехи типа «речевой хор» - при отношениях SNR > 2 дБ.

Спасибо за внимание